Cass Business School
City of London

Cass means business

Faculty of Actuarial Science and Statistics

Identification and Separation of DNA Mixtures Using Peak Area Information.

(Updated Version of Statistical Research Paper No. 25)

R.G. Cowell, S.L. Lauritzen and J. Mortera.

Statistical Research Paper No. 27

February 2006

ISBN 1-901615-94-4

Cass Business School
106 Bunhill Row
London EC1Y 8TZ
T +44 (0)20 7040 8470
www.cass.city.ac.uk

“Any opinions expressed in this paper are my/our own and not necessarily those of my/our employer or anyone else I/we have discussed them with. You must not copy this paper or quote it without my/our permission”.

Identification and Separation of DNA Mixtures using Peak Area Information

R. G. Cowell*
Faculty of Actuarial Science and Statistics,
Cass Business School,
106 Bunhill Row,
London EC1Y 8TZ, UK.

S. L. Lauritzen
Department of Statistics,
University of Oxford
1 South Parks Road,
Oxford OX1 3TG, U.K.

J. Mortera
Dipartimento di Economia
Università Roma Tre
Via Ostiense, 139
00154 Roma, Italy

February 14, 2006

THIS IS AN UPDATED VERSION OF
RESEARCH REPORT 25
SIR JOHN CASS BUSINESS SCHOOL
CITY UNIVERSITY
DATED NOV 2004

*Corresponding author: email: rgc@city.ac.uk fax: +44 (0)20 7040 8572 tel +44 (0)20 7040 8454

Abstract

We introduce a new methodology, based upon probabilistic expert systems, for analysing forensic identification problems involving DNA mixture traces using quantitative peak area information. Peak area is modelled with conditional Gaussian distributions. The expert system can be used for ascertaining whether individuals, whose profiles have been measured, have contributed to the mixture, but also to predict DNA profiles of unknown contributors by separating the mixture into its individual components. The potential of our probabilistic methodology is illustrated on case data examples and compared with alternative approaches. The advantages are that identification and separation issues can be handled in a unified way within a single probabilistic model and the uncertainty associated with the analysis is quantified. Further work, required to bring the methodology to a point where it could be applied to the routine analysis of casework, is discussed.

Some key words and phrases: DNA mixture, forensic identification, mixture separation, probabilistic expert system, peak weight.

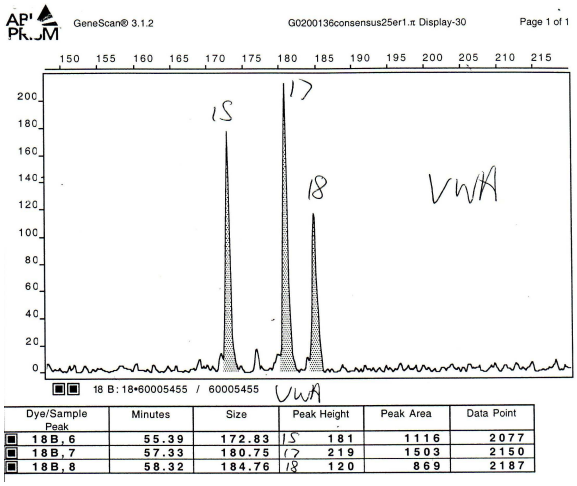
1 Introduction

Probabilistic expert systems (PES) for evaluating DNA evidence were introduced by Dawid et al. [1]. In a general review of the analysis of DNA evidence, Foreman et al. [2] include several applications of PES and emphasize their potential by predicting that this methodology "will offer solutions to DNA mixtures and many more complex problems in the future.".

This article is concerned with the analysis of *mixed traces* where several individuals may have contributed to a DNA sample left at a scene of crime. Mortera et al. [3] showed how to construct a PES using information about which alleles were present in the mixture, and we refer to this article for a general description of the problem and for genetic background information. Other earlier contributions based solely on allelic presence in the mixture are Evett et al. [4] and Weir et al. [5].

The results of a DNA analysis are usually represented as an *electropherogram* (EPG) measuring responses in *relative fluorescence units* (RFU) and the alleles in the mixture correspond to peaks with a given height and area around each allele, see Figure 1. The band intensity around each allele in the relative fluorescence units represented, for example, through their *peak areas*, contains important information about the composition of the mixture.

Experiments using heterozygous samples and mixtures prepared in known proportions have provided information on the variability of peak imbalance and the extent of stutter that can be expected when amplifying a DNA mixture. This information was used by Clayton et al. [6] to formulate general guidelines for forensic experts in order to resolve DNA mixtures based on quantitative peak area information. Gill et al. [7] built a computer program to estimate the proportion of the individual contributions in two-person mixtures and to rank the genotype combinations based on minimizing a residual sum of squares.



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
18B,6	55.39	172.83	15 181	1116	2077
18B,7	57.33	180.75	17 219	1503	2150
18B,8	58.32	184.76	18 120	869	2187

Figure 1: An electropherogram (EPG) of marker VWA from a mixture. Peaks represent alleles at 15, 17 and 18 and the areas and height of the peaks express the quantities of each. Since the peak around 17 is high, this indicates two alleles with repeat number 17. This image is supplied courtesy of LGC Limited, 2004.

More recently, Bill et al. [8] have developed PENDULUM, a computer package to automate the guidelines in [6] and [7]. First, a list of all possible genotype combinations is made and those outside heterozygous peak balance limits are eliminated; then the list is scored with respect to mixture proportion. The possibility of allelic dropout is considered, but other artifacts, such as stutter, are not accounted for. The primary purpose of PENDULUM is to eliminate unreasonable genotypic combinations. It also ranks the genotypes but this is not based on a probabilistic order, so no quantification of the uncertainty in the analysis is possible. Evidential calculations cannot be carried out directly within PENDULUM, however they may be performed by using the output of PENDULUM as the input to an external probabilistic model.

Perlin and Szabady [9] and Wang et al. [10] used numerical methods known as *Linear Mixture Analysis* (LMA) and *Least Square Deconvolution* (LSD) for separating mixture profiles using peak area information. Both methods are based on enumerating a complete set of possible genotypes that may have generated the mixture profile, on the assumptions that the mixture proportion of the contributors' DNA in the sample is constant across markers, so that the peak area of an allele will be approximately proportional to the proportion of that allele in the mixture. This may be used to calculate — via a least squares heuristic — an estimate for the mixture proportion. The major difference between the two methods is that Perlin and Szabady seek a single mixture proportion estimated using all of the markers simultaneously, whilst Wang et al. estimate a mixture proportion for each marker separately and then eliminate genotype combinations giving inconsistent estimates of this proportion across markers. Thus the methods of both [9] and [10] share features with that of Bill et al. [8].

The methods utilizing peak area information described above are not probabilistic in nature, nor do they use information about allele frequency. In contrast, the methodology proposed in Evett et al. [11] combines a model using the gene frequencies with a model describing variability in scaled peak areas to calculate likelihood ratios and study their sensitivity to assumptions about the mixture proportions.

Our approach incorporates elements similar to all of those described above, but unifies these in a single Bayesian network model. More specifically, we build a PES for mixture traces based on conditional Gaussian distributions for the peak areas, given the composition of the true DNA mixtures; see Chapter 7 of [12] as well as [13]. The exact same network is then used both for an evidential calculation as well as for the separation of DNA mixtures, with the additional benefit of a full probabilistic quantification of any uncertainty associated with the analysis.

The main focus of the present paper is to illustrate the basic ideas in a new methodology for resolving DNA mixtures based on PES. For the sake of clarity and simplicity, we only consider a DNA mixture from exactly two contributors, which seems to be the most common scenario in forensic casework [14]. We do not allow for further complications such as stutter, dropout alleles, etc. In order to develop the methodology into a practical tool for forensic laboratories these additional complications will need to be considered. However, we emphasize that the flexibility and modularity of the PES approach readily enables extension and modification of our network to include complications such as an unknown number of contributors, indirect evidence, dropout, stutter, etc. along the lines given in [3].

An analysis of a mixed trace can have different purposes, several of which can be relevant simultaneously, making a unified approach particularly suitable. However, for the sake of exposition we consider the issues separately. The first focus of our analysis will be that of *evidential calculation*, detailed in § 4. Here a *suspect* with known genotype is held and we want to determine the likelihood ratio for the hypothesis that the suspect has contributed to the mixture vs. the hypothesis that the contributor is a randomly chosen individual. We distinguish two cases: the other contributor could be a *victim* with a known genotype or a *contaminator* with an unknown genotype, possibly without a direct relation to the crime. This could be a laboratory contamination or any other source of contamination from an unknown contributor.

Another use of our network is the *separation of profiles,* i.e. identifying the genotype of each of the possibly unknown contributors to the mixture, the evidential calculation playing a secondary role. This use is illustrated in § 5.

2 Basic model assumptions

We assume the usual Mendelian genetic model for the allele composition of the mixture traces with known gene frequencies of single STR alleles, using those reported in Evett et al. [11] and Butler et al. [15] for U.S. Caucasians. We use the latter for analysing data

taken from Clayton et al. [6], Perlin and Szabady [9][1] and Wang et al. [10].

We first present a description of the model before introducing the mathematical details. In essence, the PES is a probabilistic model for relating the pre-amplification and post-amplification relative amounts of DNA in a mixture sample. The model is idealized in that it ignores complicating artifacts such as stutter, drop-out alleles and so on, and assumes that the mixture is made up of DNA from two people, who we refer to as p1 and p2. Now prior to amplification, and provided the mixture sample has not been degraded to the point of breaking up tissue cells, the sample put into the amplification apparatus will consist of an unknown number of cells from p1 and a further unknown number of cells from p2. Then, with every cell containing exactly two alleles from each marker, the fraction or proportion of cells from p1 is also a common measure across the markers of the amount of DNA from p1. We denote this common fraction, or proportion, by θ.

In an ideal amplification apparatus, during each amplification cycle the proportion of alleles of each allelic type would be preserved without error. We model departures from this ideal as random variation using the Gaussian distributions in (1), whose mean for each allele is its pre-amplification proportion for the marker system it belongs to. The variance has a simple dependence on the mean such that in the two limiting cases of (i) the pre-amplification proportion is zero, or (ii) the pre-amplification proportion is unity, the variance is zero. In the case of (i) this means that if there is no allele of a certain type in the mixture prior to amplification, there is none post-amplification. In the case of (ii) this means that if for a given marker there is only one allelic type present in the mixture pre-amplification, then only that type is present in that marker post-amplification. Our model introduces an additional variance term to represent other measurement error, represented by ω^2.

The post-amplification proportions of alleles for each marker are represented in the peak area information, which we include in the analysis through the *relative peak weight*. The (absolute) *peak weight* w_a of an allele with *repeat number* a is defined by scaling the peak area with the repeat number as

$$w_a = a\alpha_a,$$

where α_a is the peak area around allele a. Multiplying the area with the repeat number is a crude way of correcting for the fact that alleles with a high repeat number tend to be less amplified than alleles with a low repeat number. For issues concerning heterozygous imbalance see [16].

We further assume that

- The pre-amplification mixture proportion θ is constant across markers, for the reasons outlined above;
- The peak weight for an allele is approximately proportional to the amount of DNA of that allelic type;

[1]This dataset has an observed allele 25.2 of the marker FGA. As none of the 302 subjects in [15] had this allele, we chose to use 1/604=0.00166 as its frequency. For the same reason, in the example taken from [6] allele 36 of the D21 system was assigned a frequency of 0.00166.

- The peak weight for an allele possessed by both contributors is the sum of the corresponding weights for the two contributors.

To avoid arbitrariness in scaling we consider the observed *relative peak weight* r_a, obtained by scaling with the total peak weight as

$$r_a = w_a/w_+, \quad w_+ = \sum_a w_a,$$

so that then $\sum_a r_a = 1$.

Our simple model for the relative peak weight, denoted by the random variable R_a, assumes a Gaussian error distribution

$$R_a \sim \mathcal{N}(\mu_a, \tau_a^2), \quad \mu_a = \{\theta n_a^{(1)} + (1-\theta) n_a^{(2)}\}/2, \tag{1}$$

where θ is the proportion, or fraction, of DNA in the mixture originating from the first contributor, $n_a^{(i)}$ is the number of alleles with repeat number a possessed by person i.

The error variance τ_a^2 has the form

$$\tau_a^2 = \sigma^2 \mu_a (1-\mu_a) + \omega^2 \tag{2}$$

where σ^2 and ω^2 are variance factors for the contributions to the variation from the amplification and measurement processes.

The model can be seen as a second order approximation to a more sophisticated model based on gamma distributions for the absolute scaled peak weights (to be discussed elsewhere).

In addition we need to consider the correlation between weights due to the fact that they must add up to unity. If this is the only source of correlation, its inferential effect can be taken correctly into account by using the variance structure

$$\tau_a^2 = \sigma^2 \mu_a + \omega^2 \tag{3}$$

and considering the complete set of observed peak weights as observed evidence, as argued in Appendix A. Note that this is in contrast to Cowell et al. [17] who ignored the correlation without modifying the variance from (2) to (3), but essentially obtained results with the same qualititative behaviour as in the present paper.

Unless stated otherwise, we have used $\sigma^2 = 0.01$ and $\omega^2 = 0.001$, corresponding approximately to a standard deviation for the observed relative weight of about

$$\sqrt{0.01/4 + 0.001} = 0.06$$

for $\mu_a = 0.5$ substituted into (2). These parameter values imply that when amplifying DNA from one heterozygous individual (for which $\mu_a = 0.5$), an r_a value at two standard deviations from the mean would give a value of $0.38/0.62 = 0.61$ for the ratio of the minor to the major peak area; this is about the limit of variability in peak imbalance that has

been reported in the literature [18], and suggests that our chosen parameter values are perhaps conservative.

In general the variance factors may depend on the marker and on the amount of DNA analysed, but for simplicity we use the values above. Our PES model is robust to small changes in these parameter estimates. We are planning a full data analysis in order to refine the estimates and obtain a proper calibration of the variances for use in casework.

The simple model above seems in any case sufficiently accurate and adequate for the purposes of the present paper, and has the advantage that the calculations may be performed quickly using any available Bayesian network software that implements evidence propagation for conditional-Gaussian networks.

3 Bayesian networks for DNA mixtures with peak weights

3.1 Background

Here we give a very brief description of the basic ingredients of a Bayesian network or probabilistic expert system. Complete details can be found in [12]. A Bayesian network represents, by means of a *directed acyclic graph* (DAG), the complex probabilistic relationships of dependence and independence among a set of variables. (See Figure 11 for a simple example.) The *nodes* of the network represent the random variables and directed edges (arrows) connecting nodes describe the relationships among the variables. The joint probability structure of all the nodes in the network is determined by the conditional probability of each node given its graphical "parents"[2]. The joint probability model of all the nodes is thus expressed in terms of simple submodels. Fast and efficient computational algorithms exist for the exact calculation of marginal and conditional probabilities for the conditional-Gaussian networks used in this paper [13]. These enable the evidence (or information) on a set of nodes to be propagated to all nodes in the network and thus obtain the updated posterior conditional probabilities for all the variables represented.

Bayesian networks can easily be implemented using readily available software such as HUGIN[3]. The graphical interface can be used to specify the qualitative relationships between the variables, their values and the conditional probabilities. The network is then compiled (giving the marginal distribution of all nodes) and after evidence is inserted and propagated throughout the network the updated conditional probability distributions can be read off the nodes of interest.

The Bayesian networks constructed for the examples in this paper were implemented in HUGIN, as described in Appendix B, and also in a separate program MAIES, described in Appendix C. One reason for this duplication was to ensure correctness and additional flexibility in the specifications of the Bayesian networks. The calculations performed at a

[2]We say that node A is a graphical parent of node B if there is an edge directed from A to B.

[3]See `www.hugin.com`

high level of discretization were initially performed using MAIES, and then checked using HUGIN by exporting from MAIES non object-oriented Bayesian networks to files in the HUGIN format.

3.2 Object-Oriented Networks

Object-oriented Bayesian networks [19, 20] have a hierarchical structure where any node itself can represent a (object-oriented) network containing several *instances* of other generic *classes* of networks. This framework is particularly suited for an application area such as the present because we can exploit the similarity between elements of the networks in a modular and flexible construction, making the networks more and more complex by simply adding new objects which perform different tasks. Two recent examples of object-oriented Bayesian networks applied to forensic DNA problems are [21] and [22].

Instances have interface *input* and *output* nodes as well as ordinary nodes. Instances of a particular class have identical conditional probability tables for non-input nodes. Instances are connected by arrows from output nodes to input nodes. These arrows represent identity links whereas arrows between ordinary nodes represent probabilistic dependence. Implementation of object-oriented Bayesian networks is supported by the program HUGIN 6.4, which we use in our analyses. A more detailed description of the component object-oriented networks used in this paper may be found in Appendix B.

3.3 Maies: A PES for analysing mixed traces

As indicated above, in parallel to the development of the object-oriented networks a separate computer program called MAIES—Mixture Analysis In Expert Systems—was developed to provide an independent check of the calculations. It also provided a flexible environment for specification of input and output of data that allowed for sensitivity analysis and, for example, to provide the data in a useful form for producing posteriors plots.

The input to MAIES is simply the measured peak area information on up to four alleles per marker, the population gene frequencies of these alleles, and additional genotypic information (if available) about the potential contributors.

After entering peak area information and available genetic profiles on people, the software constructs a single Bayesian network on which the probability calculations are performed (see Figure 20). In constructing the Bayesian network the user is presented with the options of changing default values for the scale σ^2 of the amplification error variance, the measurement error variance ω^2, and the number of discrete states used for the node that represents θ, the true pre-amplification fraction of DNA originating from individual 1. Sensitivity analysis may be performed in a simple, straightforward manner by varying these three inputs. Peak areas are automatically converted to normalized weights by the program, and entered as evidence in the relevant nodes.

The user can temporarily retract or reinstate evidence on the two potential contributors to the mixture by use of menu selections, thus allowing an evidential calculation to be converted to one of deconvoluting a mixture arising from two unknown contributors, or

vice versa. A more detailed description of the networks generated by MAIES may be found in Appendix C.

4 Evidence calculations

This section illustrates (through the analysis of real mixture examples) how to use our PES to calculate the weight of the evidence—in the form of a likelihood ratio—for a given *suspect* to have contributed to a trace under different circumstances.

The evidence could consist of DNA profiles extracted from a *suspect, s,* a *victim, v,* and the mixed trace. In this case we compute the likelihood ratio in favour of the hypothesis that the victim and suspect contributed to the mixture: $H_0 : v\&s$, vs. the hypothesis that the victim and an unknown individual, u contributed to the mixture: $H_1 : v\&u$.

A variant has an *unknown contaminator*, u instead of a victim, in which case the hypotheses are $H_0 : u\&s$ versus $H_1 : 2u$.

In the results shown below (and also for the examples in § 5) the variable describing the mixture proportion θ has been discretized to having 101 states $0, 0.01, 0.02, \ldots, 0.99, 1$, but experiments indicate very low sensitivity to the discretization as long as it is not far too rough and 10-20 states would probably be fully appropriate.

4.1 Genotype of suspect and victim available

This example is taken from Wang et al. [10], stating P. Graham of the Texas Department of Public Safety as the data source. Table 1 displays the alleles observed in the mixture, the measured peak area and the relative weight on 9 markers, together with the genotypes of two potential contributors, here named suspect, s and victim, v. We will in the following refer to this data as the *Graham* data.

The evidence in this table is now entered and propagated throughout the network yielding the marginal posterior probabilities or densities of the quantities of interest. The evidence on allelic repeat number is inserted in the appropriate nodes; details on how this is done are given in Appendix B.6 and Appendix C.4. When peak area information is also used, the nodes representing the observed relative peak weights are set to their corresponding values, as illustrated in Appendix B.7 and Appendix C.5. Taking appropriate ratios in the posterior probabilities associated with the target node yields the likelihood ratio in favour of $H_0 : v\&s$ versus $H_1 : v\&u$. Table 2, column "Areas" displays the logarithm of this likelihood ratio, and column "Alleles" the corresponding ratio when only the evidence on the repeat number of the alleles is used. The last columns of Table 2 show the log-likelihood ratio when the mixture proportion θ is assumed known at given values. The same network is used to compute all the quantities in Table 2.

Note that the likelihood ratio is essentially constant in the region $0.3 < \theta < 0.4$ which is the plausible region in the light of the data. The posterior distribution of the mixture proportion θ is displayed in Figure 2. Note that this posterior distribution has its maximum around 0.34, close to the value reported in [10].

Table 1: *Graham* data showing mixture composition, peak areas, relative weights, suspect's and victim's profiles.

Marker	Alleles	Peak area	Relative weight	Suspect	Victim
D3	15	1242	0.3361		15
	16	657	0.1897	16	
	17	1546	0.4742	17	17
D5	7	486	0.0999	7	
	12	512	0.1804	12	
	13	1886	0.7198		13
D7	10	614	0.3232	10	
	11	1169	0.6768		11
D8	12	1842	0.6166		12
	13	490	0.1777	13	
	16	461	0.2057	16	
D13	8	734	0.3128		8
	9	1068	0.5120	9	9
	11	299	0.1752	11	
D18	12	440	0.1724	12	
	13	1503	0.6380		13
	15	387	0.1896	15	
D21	30	842	0.3087		30
	30.2	490	0.1808	30.2	
	31.2	509	0.1941	31.2	
	32.2	804	0.3164		32.2
FGA	22	850	0.3483		22
	23	468	0.2005	23	
	24	681	0.3045		24
	25	315	0.1467	25	
VWA	16	616	0.1900	16	
	17	2021	0.6625		17
	18	425	0.1475	18	

Table 2: Logarithm of the likelihood ratios in favour of $H_0 : v\&s$ vs. $H_1 : v\&u$ for the *Graham* data.

	Areas	Alleles	Assumed known mixture proportion								
θ			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Log_{10}LR	14.47	12.93	10.97	13.44	14.46	14.42	12.42	8.58	2.76	-7.19	-27.79

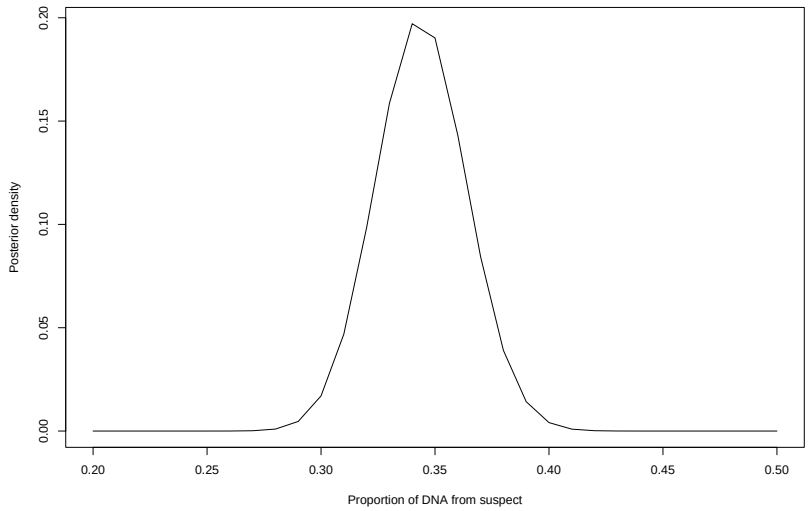


Figure 2: Posterior distribution of the mixture proportion for the *Graham* data using both the suspect's and the victim's genotype.

The inclusion of area information is indeed strengthening the evidence against the suspect, increasing the logarithm of the likelihood ratio from 12.93 to 14.47, approximately corresponding to a factor 36. This is a modest increase and reflects the fact that when information about the genotype of the victim is available, peak area does not make much difference to the likelihood ratio as the genotypes themselves are very informative.

4.2 Only genotype of suspect available

Our next example is taken from Evett et al. [11] and has only information of the genotype from one potential contributor, here named the *suspect*, whereas the other unknown contributor is termed *contaminator*. The data refers to a 10:1 mixture of two individuals. The data is displayed in Table 3 and is henceforth referred to as the *Evett* data. Table 4 displays the logarithm of this likelihood ratio together with the corresponding ratio when peak weights are ignored, and the ratios when the mixture proportion θ is assumed known at given values.

Note that the strengthening of evidence against the suspect is more dramatic when information on the contaminator is absent: the logarithm of the likelihood ratio changes from 4.4 to 8.23, corresponding to an additional factor around 6000, as compared to a factor 36 above.

Also here the likelihood ratio is essentially constant over a region which completely covers the posterior plausible range $0.85 < \theta < 0.95$.

The posterior distribution of the mixture proportion θ is displayed in Figure 3. The maximum occurs around the value 0.90 which is a little off the true 10:1 mixture proportion.

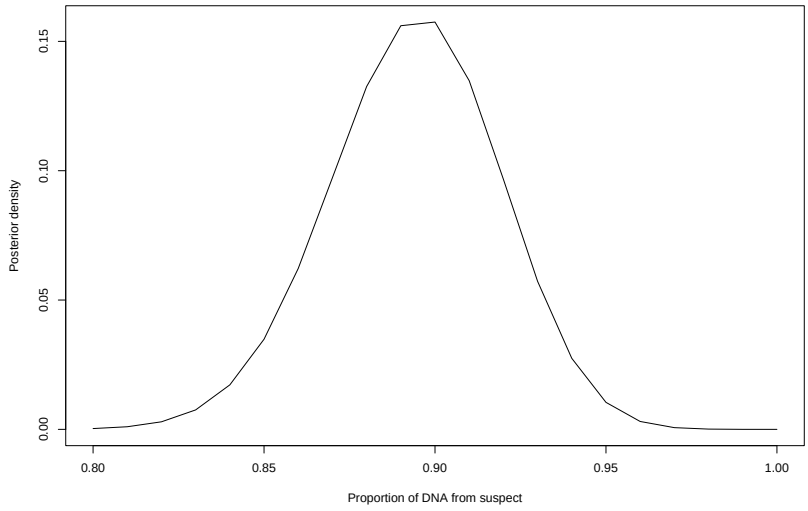


Figure 3: Posterior distribution of the mixture proportion for the *Evett* data using the suspect's genotype.

Table 3: *Evett* data showing mixture composition, peak areas and relative weights from a 10:1 mixture of two individuals, with suspect's genotype specified.

Marker	Alleles	Peak area	Relative weight	Suspect
D8	10	6416	0.4347	10
	11	383	0.0285	
	14	5659	0.5368	14
D18	13	38985	0.8871	13
	16	1914	0.0536	
	17	1991	0.0592	
D21	59	1226	0.0525	
	65	1434	0.0676	
	67	8816	0.4284	67
	70	8894	0.4515	70
FGA	21	16099	0.5699	21
	22	10538	0.3908	22
	23	1014	0.0393	
THO1	8	17441	0.4015	8
	9.3	22368	0.5985	9.3
VWA	16	4669	0.4170	16
	17	931	0.0884	
	18	4724	0.4747	18
	19	188	0.0199	

Table 4: Logarithm of the likelihood ratios in favour of $H_0 : u\&s$ vs. $H_1 : 2u$ for the *Evett* data.

	Areas	Alleles	Assumed known mixture proportion								
θ			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Log_{10}LR	8.23	4.40	-237.47	-138.73	-77.53	-31.68	5.06	8.32	8.52	8.53	8.53

The absolute value of the likelihood ratios are slightly different from those given by [11], who report a logarithm of the likelihood ratio of 7.3. This discrepancy is most likely due to slight differences between our model and the model used by [11]. On the other hand, they report a likelihood ratio based on allele presence alone of 5800, whereas we find a ratio around 25000 using the gene frequencies reported in their paper, and insist the latter must be the correct value.

5 Separation of mixtures

Deconvolution of mixtures or separating a mixed DNA profile into its components has been studied by Perlin and Szabady [9], Wang et al. [10], and Bill et al. [8], among others. Here, we show how separation of mixtures can be solved by the same network model used for evidence calculations. A mixed DNA profile has been collected and the genotypes of one or more unknown individuals who have contributed to the mixture is desired, for example with the purpose of searching for a potential perpetrator among an existing database of DNA profiles.

For a two-person mixture, the easiest case to consider is clearly that of separation of a single unknown profile, i.e. when the genotype of one of the contributors to the mixture is known. The case when both contributors are unknown is more difficult. In the latter situation this is only possible to a reasonable accuracy when the contributions to the DNA mixture has taken place in quite different proportions.

We have chosen to show two alternative methods for predicting the genotype of the unknown contributor(s). In the first method we report the most probable genotype (or pair of genotypes) of the unknown contributor(s) for each marker separately. This result is obtained directly from the standard propagation method in the probabilistic expert system, known as sum-propagation. Note that this genotype is not necessarily the jointly most probable across markers. We therefore also report the joint probability of the genotypes chosen in this way. If this happens to be larger than 0.5, the most probable genotype has clearly been identified.

The second method calculates, by a method termed *semimax*-propagation, the most likely joint configuration of all unobserved discrete nodes, given the evidence available, and reports the genotypes of the unknown contributor(s) associated with this configuration. The semimax propagation first integrates over all unobserved continuous variables and then performs max-propagation as described in [12], Section 6.4.1, to identify the most probable configuration. Note again, that this may not be the most probable genotype across markers. There is no general efficient method for calculating the latter, but identifying the two configurations above and reporting their joint probabilities would be fully satisfactory for most purposes as they are most interesting when their joint probability is high.

The two methods generally give results that agree quite closely, the difference largely being due to correlations between the markers originating from the fact that the fraction of DNA supplied by each contributor is unknown. When this fraction is well determined by the evidence, the markers are close to being independent. In such cases the two methods

tend to give identical results. It then also holds that the joint posterior probability of the genotypes of the unknown contributors is approximately equal to the product of those probabilities for each marker separately.

It would seem appropriate to report a list of probable genotypes for the unknown contributor(s), with their associated probabilities, but this would demand a slightly more sophisticated calculation and is beyond the scope of this particular paper.

5.1 Separating a single unknown profile

Our next example uses data from Perlin and Szabady [9], henceforth referred to as the *Perlin* data, displayed in Table 5.

The two individuals contributing to the mixture are here named *suspect* and *victim* and Table 6 displays the predicted genotype of the suspect, using information from the victim alone.

As in [9] the genotype of the unknown contributor is essentially determined exactly and the posterior distribution of the mixture proportion concentrates around the true value of 0.7, as displayed in Figure 4.

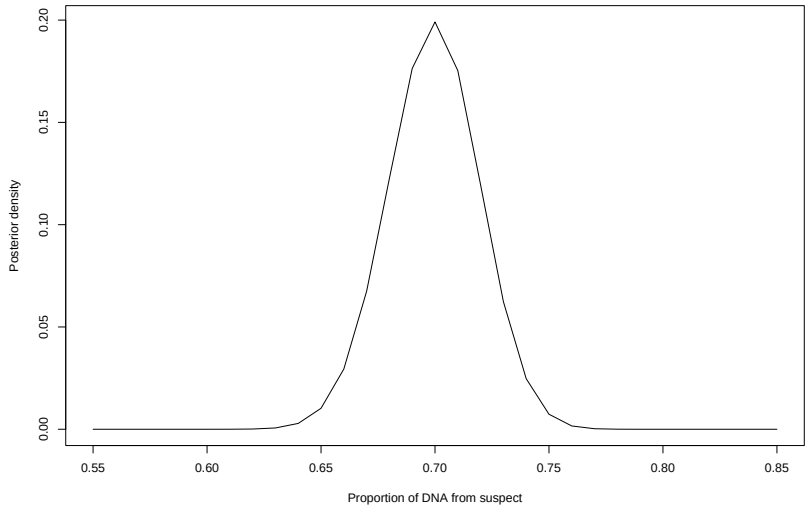


Figure 4: Posterior distribution of the mixture proportion for the *Perlin* data, using genotypic information on the victim only.

For comparison we have also made a similar calculation for the other two examples. The results are displayed in Table 7 and Table 8.

Table 5: *Perlin* data showing mixture composition, peak areas, relative weights, suspect's and victim's genotypes from a 7:3 mixture of two individuals.

Marker	Alleles	Peak area	Relative Weight	Suspect	Victim
D2	16	0.3190	0.1339		16
	18	0.6339	0.2992	18	
	20	0.3713	0.1947		20
	21	0.6758	0.3722	21	
D3	14	1.0365	0.5010	14	14
	15	0.9635	0.4990	15	15
D8	9	0.7279	0.2832	9	
	12	0.2749	0.1426		12
	13	0.6813	0.3829	13	
	14	0.3160	0.1913		14
D16	11	1.4452	0.6801	11	
	13	0.2889	0.1607		13
	14	0.2660	0.1593		14
D18	12	0.3443	0.1504		12
	13	0.6952	0.3290	13	
	14	0.6755	0.3443	14	
	17	0.2850	0.1764		17
D19	12.2	0.6991	0.3109	12.2	
	14	0.6060	0.3092		14
	15	0.6949	0.3799	15	
D21	27	0.2787	0.1289		27
	29	0.7876	0.3913	29	
	30	0.9337	0.4798	30	30
FGA	19	1.0580	0.4621	19	19
	24	0.2830	0.1561		24
	25.2	0.6589	0.3817	25.2	
THO1	6	0.3178	0.1268		6
	7	1.0074	0.4691	7	7
	9	0.6749	0.4041	9	
VWA	17	1.4755	0.7265	17	
	18	0.5245	0.2735		18

Table 6: Predicted genotype of suspect for the *Perlin* data, using genotype information for victim only. All markers are correctly identified by both sum and semi-max propagation.

Marker	Genotype	Probability
D2	18 21	1
D3	14 15	1
D8	9 13	1
D16	11 11	1
D18	13 14	1
D19	12.2 15	1
D21	29 30	1
FGA	19 25.2	1
THO1	7 9	1
VWA	17 17	1

Table 7: Predicted genotype of suspect for *Graham* data, using genotype for victim only. All markers are correctly identified by both sum and semi-max propagation. The number in brackets is the product of individual marker probabilities.

Marker	Genotype	Probability
D3	16 17	0.982624
D5	7 12	1
D7	10 10	0.984372
D8	13 16	1
D13	9 11	0.995181
D18	12 15	1
D21	30.2 31.2	1
FGA	23 25	1
VWA	16 18	1
joint	0.962504	(0.962606)

Table 8: Predicted genotype of contaminator for *Evett* data, using information from suspect. Identical results are obtained using sum and semi-max propagation. The number in brackets is the product of individual marker probabilities.

Marker	Genotype	Probability
D8	11 14	0.834050
D18	16 17	1
D21	59 65	1
FGA	21 23	0.815391
THO1	9.3 9.3	0.826110
VWA	17 19	1
joint	0.567333	(0.561818)

The situation for the *Graham* data is similar to the *Perlin* data: all markers are correctly identified, with probabilities very close to 1 in all cases. Analysis of the *Evett* data yield probabilities between 0.8 and 1 on all markers. Evett et al. [11] does not contain the genotype of the second contributor so we do not know whether there are classification errors for this example. Figure 3 and Figure 5 display the posterior distribution of the mixture proportion for these two cases.

5.2 Separating two unknown profiles

We now turn to the problem of separating a mixture into two components, using peak area and repeat number information but no information regarding the two contributors to the mixture. Using only this information will lead to an identifiability problem in assigning genotype combinations to each person, because of the symmetry between the individuals p1 and p2 in the network of Figure 20 or in the equivalent object-oriented network Figure 18.

To remove this problem it is sufficient to enter evidence that the pre-amplification proportion of DNA in the sample from individual p1 is at least one half of the total DNA in the sample. (The alternative, that individual p1 contributes at most half of the DNA to the mixture sample could as equally well be used to break the symmetry.) Using HUGIN this symmetry breaking may be achieved by entering likelihood evidence directly into the fraction node; in MAIES direct entering of likelihood evidence is not possible, so instead this is achieved by entering evidence on the `sym` node. The node `sym` has two possible states, $\theta \geq 0.5$ and $\theta \leq 0.5$. Selecting one state as evidence breaks the symmetry (the user does this via a menu selection). It is important to note that setting the proportion of DNA originating from p1 to be less that 0.5 in the pre-amplification mixture does not mean that post-amplification the proportion of DNA originating from p1 is necessarily also less than 0.5, the variance structures in our model can allow this to be greater than 0.5. What it

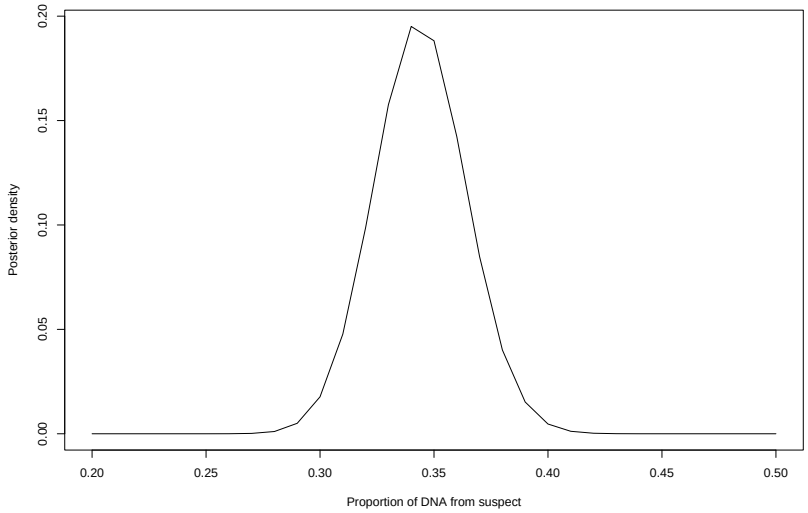


Figure 5: Posterior distribution of mixture proportion for *Graham* data using genotypic information from the victim only.

does imply is that the posterior distribution of the pre-amplification fraction will be zero for values greater than 0.5.

Our first example uses the *Evett* data, ignoring the information on the suspect. The posterior distribution of the mixture proportion θ is displayed as the solid curve in Figure 6. The distribution is similar in shape to that in Figure 3, which uses the suspect genotype information. The broken curve in Figure 6 shows the posterior using the larger variance factor $\sigma^2 = 0.1$. We note that this change of variance by an order of magnitude has a notable effect on the posterior distribution of mixture proportion.

The predicted genotypes of the two contributors are shown in Table 9, with the suspect's profile being predicted correctly for both choices of variance even though the probability of the chosen genotype is strongly reduced when the larger variance factor is used. Note that the probabilities in the left half of Table 9 are the same as those in Table 8 (to the accuracy given). This would not normally be expected, but for this example it turns out that in separating the profiles the genotype of person 1 is predicted with high certainty to be the same as the suspect. Hence adding in the suspect's profile as was done for the calculations of Table 8 would have very little affect on the predictions made by the system for the genotype of the contaminator.

Our next example uses the *Perlin* data. The posterior distribution for θ is shown as the solid curve in Figure 7, with the mode at 0.69 very close to the value reported of 0.7. The predicted genotypes of the two contributors are shown in Table 10, with all but one of the classifications correct. The sum of the joint probabilities for the two chosen genotypic combinations is around 0.61, indicating that other plausible explanations are available. This is essentially due to uncertainty about the genotype for marker VWA and to a lesser

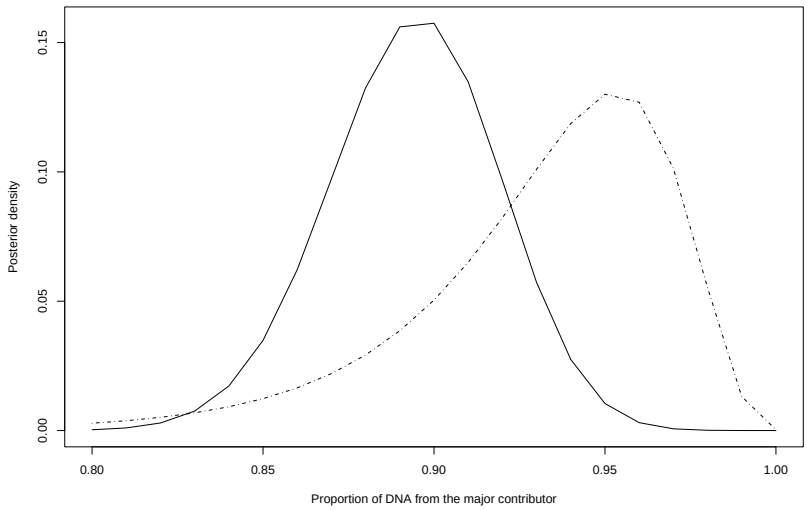


Figure 6: Posterior distribution of mixture proportion from *Evett* data using no genotypic information: solid line $\sigma^2 = 0.01$, broken line $\sigma^2 = 0.1$.

Table 9: Predicted genotypes of both contributors for *Evett* data with $\sigma^2 = 0.01$ and $\sigma^2 = 0.1$. Identical results are obtained using sum and semi-max propagation, with suspect (p1) correct on every marker. The number in brackets is the product of individual marker probabilities.

	$\sigma^2 = 0.01$			$\sigma^2 = 0.1$		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D8	10 14	11 14	0.834050	10 14	11 14	0.654367
D18	13 13	16 17	1	13 13	16 17	0.876868
D21	67 70	59 65	1	67 70	59 65	0.999405
FGA	21 22	21 23	0.815391	21 22	21 23	0.489847
THO1	8 9.3	9.3 9.3	0.826110	8 9.3	9.3 9.3	0.574267
VWA	16 18	17 19	1	16 18	17 19	0.999390
joint	0.567333		(0.561818)	0.161705		(0.161215)

extent for the marker D19.

Increasing σ^2 by a factor of 10 to $\sigma^2 = 0.1$ yields the posterior distribution for θ shown by the broken line Figure 7. In this case the effect of choosing an inflated variance factor is dramatic, also yielding reduced genotype probabilities and several classification errors as shown in Table 11. Note also that here there is a marked discrepancy between probability of the joint genotype and the product of the probabilities for each marker.

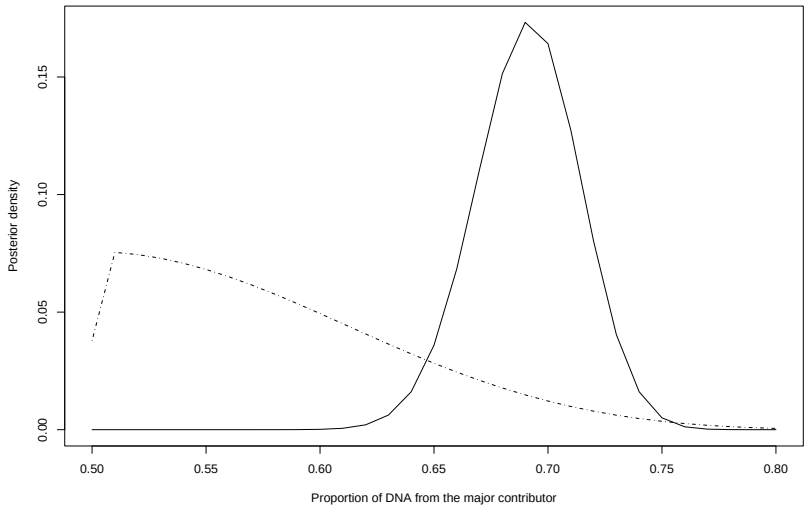


Figure 7: Posterior distribution of mixture proportion from *Perlin* data using no genotypic information: solid line $\sigma^2 = 0.01$, broken line $\sigma^2 = 0.1$.

Similar behaviour occurs in our third example that uses the *Graham* data. The posterior distribution of θ is shown as the solid curve in Figure 8, with a maximum for the major contributor around 0.65; the predicted profiles are shown in Table 12, with one classification error. However note for this classification error (in D7, using sum-propagation) the probability assigned to the genotype pair is around 0.66, with the correct classification (picked out by the semi-max method) having a probability of around 0.33. Note that the two chosen genotypes together account for essentially all of the probability mass.

Increasing the variance factor σ^2 to 0.1 yields more classification errors but is also accompanied by much lower probabilities, as shown in Table 13. The corresponding posterior distribution of θ is plotted as the broken line in Figure 8.

Table 10: Predicted genotypes of both contributors for *Perlin* data with $\sigma^2 = 0.01$. The number in brackets is the product of individual marker probabilities. For semi-max propagation all classifications are correct but for sum propagation there is a classification error in marker VWA (italicized).

	sum prop			semi-max		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D2	18 21	16 20	0.996545	18 21	16 20	0.996545
D3	14 15	14 15	0.974334	14 15	14 15	0.974334
D8	9 13	12 14	0.992179	9 13	12 14	0.992179
D16	11 11	13 14	0.994388	11 11	13 14	0.994388
D18	13 14	12 17	0.999520	13 14	12 17	0.999520
D19	12.2 15	14 14	0.796869	12.2 15	14 14	0.796869
D21	29 30	27 30	0.955125	29 30	27 30	0.955125
FGA	19 25.2	19 24	0.971191	19 25.2	19 24	0.971191
THO1	7 9	6 7	0.922004	7 9	6 7	0.922004
VWA	*17 18*	*17 17*	0.549705	17 17	18 18	0.393374
joint	0.353239		(0.358721)	0.261764		(0.256704)

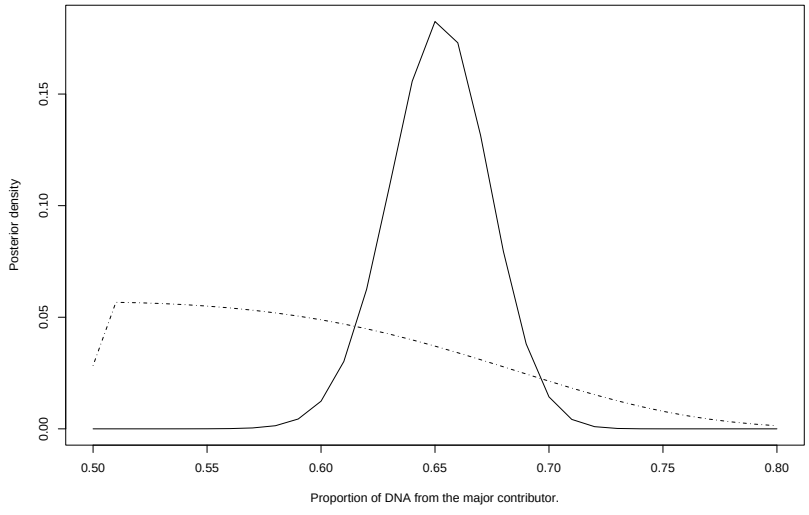


Figure 8: Posterior distribution of mixture proportion from *Graham* data using no genotypic information: solid line $\sigma^2 = 0.01$, broken line $\sigma^2 = 0.1$.

Table 11: Predicted genotypes of both contributors for *Perlin* data with $\sigma^2 = 0.1$. The number in brackets is the product of individual marker probabilities. There are classification errors in five markers (italicized).

	sum prop			semi-max		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D2	18 21	16 20	0.285179	18 21	16 20	0.285179
D3	14 15	14 15	0.461512	14 15	*15 15*	0.229006
D8	9 13	12 14	0.279718	9 13	12 14	0.279718
D16	*11 13*	*11 14*	0.315883	11 11	13 14	0.270330
D18	13 14	12 17	0.291047	13 14	12 17	0.291047
D19	*14 15*	*12.2 14*	0.218410	12.2 15	14 14	0.170503
D21	29 30	27 30	0.357621	29 30	27 30	0.357621
FGA	19 25.2	19 24	0.324954	19 25.2	*24 24*	0.129379
THO1	7 9	6 7	0.322619	7 9	6 7	0.322619
VWA	*17 18*	*17 17*	0.364098	*17 18*	*17 17*	0.364098
joint	2.6978e-05		(1.0091e-05)	3.1643e-05		(1.332e-06)

Table 12: Prediction of two unknown genotypes for *Graham* data, with $\sigma^2 = 0.01$. The number in brackets is the product of individual marker probabilities. There is a classification error in marker D7 (italicized).

	sum prop			semi-max		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D3	16 17	15 17	0.963136	16 17	15 17	0.963136
D5	7 12	13 13	0.994966	7 12	13 13	0.994966
D7	*11 11*	*10 11*	0.659653	10 10	11 11	0.326069
D8	13 16	12 12	0.835225	13 16	12 12	0.835225
D13	9 11	8 9	0.981719	9 11	8 9	0.981719
D18	12 15	13 13	0.931912	12 15	13 13	0.931912
D21	30.2 31.2	30 32.2	0.979851	30.2 31.2	30 32.2	0.979851
FGA	23 25	22 24	0.985227	23 25	22 24	0.985227
VWA	16 18	17 17	0.967872	16 18	17 17	0.967872
joint	0.451047		(0.451327)	0.227689		(0.223093)

Table 13: Prediction of two unknown genotypes for *Graham* data, using $\sigma^2 = 0.1$. There are now classification errors in six markers (italicized).

	sum prop			semi-max		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D3	16 17	15 17	0.263132	16 17	15 17	0.263132
D5	*7 13*	*12 13*	0.311064	7 12	13 13	0.298917
D7	*10 11*	*10 11*	0.335654	10 10	11 11	0.133020
D8	*12 13*	*12 16*	0.310974	13 16	12 12	0.216863
D13	9 11	8 9	0.215735	*11 11*	8 9	0.133398
D18	*12 13*	*13 15*	0.300377	12 15	13 13	0.231450
D21	30.2 31.2	30 32.2	0.278259	30.2 31.2	30 32.2	0.278259
FGA	23 25	22 24	0.302807	23 25	22 24	0.302807
VWA	*17 18*	*16 17*	0.321188	16 18	17 17	0.262347
joint	1.7666e-05		(1.4983e-05)	6.9669e-06		(1.5486e-06)

5.3 An example using amelogenin

Our final example is taken from Appendix B of Clayton et al. [6] and illustrates the importance of the *amelogenin* marker in the analysis of DNA mixtures when the individual contributors are of opposite sex.

Peak area analysis of the amelogenin marker in DNA recovered from a condom used in a rape attack indicated an approximate 2:1 ratio for the amount of female to male DNA contributing to the mixture. Peak area information was available on six other markers, the information is shown in Table 14; we shall refer to this as the *Clayton* data.

In Table 15 we show the results of separating the mixture using peak area information only, without using information on the victim. All markers are correctly identified. Note in particular that the genotypes for the marker THO are identified correctly. Clayton et al. [6] were only able to do this after the victim's profile was taken into account, because without this information the alternative genotype combination $(\{7,7\},\{5,5\})$ could also have explained the observed peak areas with an approximate 2:1 imbalance in the contributors' DNA. In our analysis we estimate that this alternative combination is around 258 times less likely than the correct designation.

Figure 9 shows the posterior distribution of the mixture proportion; the peak at around 0.65 corresponds to a mixture ratio of 1.86:1, in line with the approximate 2:1 estimated in [6].

Table 14: *Clayton* data showing mixture composition, peak areas and relative weights together with the DNA profiles of both victim and suspect. For the marker D21 the allele designation in brackets is as given in [6] using the Urquhart et al. [23] labelling convention

Marker	Alleles	Peak area	Relative weight	Suspect	Victim
Amelogenin	X	1277	0.8298	X	XX
	Y	262	0.1702	Y	
D8	13	3234	0.6372		13
	14	752	0.1596	14	
	15	894	0.2032	15	
D18	14	1339	0.1462	14	
	15	1465	0.1714	15	
	16	2895	0.3612		16
	18	2288	0.3212		18
D21	28 (61)	373	0.1719	28	
	30 (65)	590	0.2913		30
	32.2 (70)	615	0.3259		32.2
	36 (77)	356	0.2109	36	
FGA	22	534	0.1547	22	
	23	2792	0.8453	23	23
THO	5	5735	0.2756		5
	7	10769	0.7244	7	7
VWA	15	1247	0.1633	15	
	16	1193	0.1667	16	
	17	2279	0.3383		17
	19	2000	0.3318		19

Table 15: Predicted genotypes of both contributors for *Clayton* data with $\sigma^2 = 0.01$. Identical results are obtained using sum and semi-max propagation, with victim (p1) and male suspect (p2) correct on every marker. The number in brackets is the product of individual marker probabilities.

	$\sigma^2 = 0.01$		
Marker	Genotype p1	Genotype p2	Probability
Amelogenin	X X	X Y	0.983115
D8	13 13	14 15	0.903013
D18	16 18	14 15	0.993166
D21	30 32.2	28 36	0.945235
FGA	23 23	22 23	0.989090
THO	5 7	7 7	0.845031
VWA	17 19	15 16	0.992738
joint	0.701988		(0.691517)

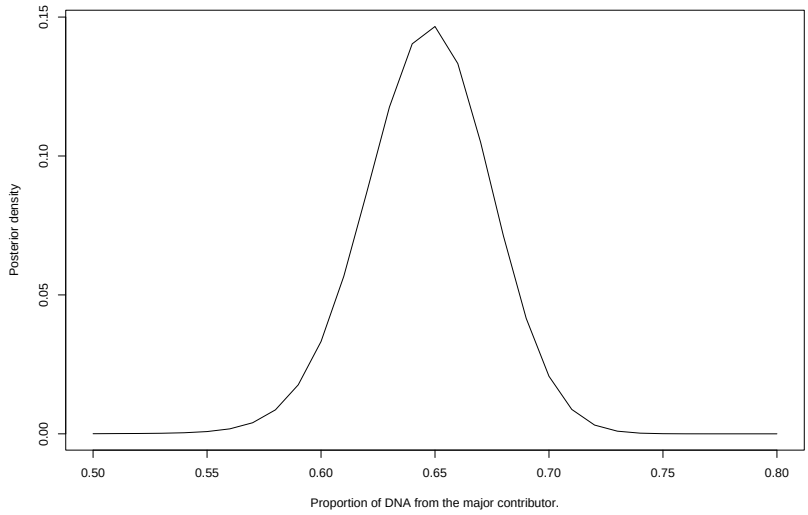


Figure 9: Posterior distribution of mixture proportion from *Clayton* data using no genotypic information, with $\sigma^2 = 0.01$.

6 Discussion

In the previous sections we have demonstrated how a probabilistic expert system can be used for analysing DNA mixtures using peak area information, yielding a coherent way of predicting genotypes of unknown contributors and assessing evidence for particular individuals having contributed to the mixture. The advantages of a probabilistic model-based approach over numerical separation techniques such as Linear Mixture Analysis (LMA) [9] and Least Square Deconvolution (LSD) [10] are that there is a natural and directly interpretable quantification of all uncertainties associated with the analysis; in particular, the posterior distribution of the mixture proportion can be computed. Furthermore, the analysis is extendable to similar but different situations using the modularity and flexibility of the PES approach. This includes complications such as more than two potential contributors, multiple traces, indirect genotypic evidence, stutter, etc.

The examples considered have also demonstrated that there are issues which need further consideration. In particular it appears that the performance of the system is sensitive to large changes in the scaling factors we used to model the variation in the amplification and measurement processes. This is a serious problem which needs attention. Preliminary investigations seem to indicate that the variance factor depends critically on the total *amount* of DNA available for analysis. As this necessarily is varying from case to case, a calibration study should be performed to take this properly into account. In any case we find it comforting that the system itself would warn against trusting an uncertain prediction, by yielding an associated low classification probability.

Methods for diagnostic checking and validation of the model should be developed based upon comparing observed weights to those predicted when genotypes are assumed correct. Such methods could also be useful for calibrating the variance parameters σ^2 and ω^2. To indicate a possible way ahead we note that the network can itself be used for predicting peak weight given a hypothesized composition of the mixture and of the two contributors. Table 16 gives the predicted peak weights for the *Perlin* data based on the repeat numbers in the mixture composition, the true mixture composition, and on the suspect's and victim's genotype. The last two columns show the limits of the 95% predictive interval $[\mu_a - 1.96\tau, \mu_a + 1.96\tau]$ for the weight. For this purpose we use the variance structure (2) as only the marginal distribution of the peak weights are involved so the correlations do not interfere. For a 95% predictive interval we might expect about one of the weights of the table to lie outside of its predicted interval, as about 21 of the 31 intervals are independent (the weights for each marker must add to one); all expected weights are within their intervals, indicating that the variance is not too small.

The predicted peak weights are also useful for identifying measurement errors. For example, if the predicted weight is of the same order of magnitude as the cut-off threshold, the peak is likely to be missed. In future work when we incorporate artifacts this will be especially useful for analysing mixtures with low copy number to distinguish noise from signal.

Another issue to be further investigated is the possibility of using a model based on gamma distributed absolute peak weights, avoiding the somewhat unfortunate fact that

Table 16: Prediction of relative peak weight for *Perlin* data, using the mixture, the suspect's and the victim's DNA composition.

Marker	Allele	Relative Weight	Predicted relative weight	
			$\mu_a - 1.96\tau$	$\mu_a + 1.96\tau$
D2	16	0.1339	0.0565	0.2435
	18	0.2992	0.2378	0.4622
	20	0.1947	0.0565	0.2435
	21	0.3722	0.2378	0.4622
D3	14	0.5010	0.3840	0.6160
	15	0.4990	0.3840	0.6160
D8	9	0.2832	0.2378	0.4622
	12	0.1426	0.0565	0.2435
	13	0.3829	0.2378	0.4622
	14	0.1913	0.0565	0.2435
D16	11	0.6801	0.5909	0.8091
	13	0.1607	0.0565	0.2435
	14	0.1593	0.0565	0.2435
D18	12	0.1504	0.0565	0.2435
	13	0.3290	0.2378	0.4622
	14	0.3443	0.2378	0.4622
	17	0.1764	0.0565	0.2435
D19	12.2	0.3109	0.2378	0.4622
	14	0.3092	0.1909	0.4091
	15	0.3799	0.2378	0.4622
D21	27	0.1289	0.0565	0.2435
	29	0.3913	0.2378	0.4622
	30	0.4798	0.3840	0.6160
FGA	19	0.4621	0.3840	0.6160
	24	0.1561	0.0565	0.2435
	25.2	0.3817	0.2378	0.4622
THO1	6	0.1268	0.0565	0.2435
	7	0.4691	0.3840	0.6160
	9	0.4041	0.2378	0.4622
VWA	17	0.7265	0.5909	0.8091
	18	0.2735	0.1909	0.4091

Gaussian distributions can take negative values. Ideally the method should be generalized to deal with higher complexity such as the simultaneous analysis of several traces, an unknown but large number of contributors, etc., and we have not as yet made a proper investigation of the computational complexity issues associated.

We also will explore how to extend the model to handle Y-chromosome and mitochondrial DNA haplotype data. Finally, we emphasize that for the moment we have not dealt with incorporating artifacts such as stutter, pull-up, allelic dropout, etc., but we hope to pursue this and other aspects in the future. It may be that in incorporating such artifacts our networks will become too complex for exact inference based on evidence propagation in Bayesian networks, and that a Monte-Carlo simulation approach may be required.

Acknowledgement

This research was supported by a Research Interchange Grant from the Leverhulme Trust. We are indebted to participants in the above grant and to Sue Pope and Niels Morling for constructive discussions. We thank Caryn Saunders for supplying the EPG image used in Figure 1. We also thank the associate editor and the referees for helpful comments.

References

[1] A. P. Dawid, J. Mortera, V. L. Pascali, and D. W. van Boxel. Probabilistic expert systems for forensic inference from genetic markers. *Scand. J. Stat.*, 29 (2002) 577–595.

[2] L. A. Foreman, C. Champod, I. W. Evett, J.A. Lambert, and S. Pope. Interpreting DNA Evidence: A Review. *Internat. Statist. Rev.*, 71 (2003) 473–495.

[3] J. Mortera, A. P. Dawid, and S. L. Lauritzen. Probabilistic expert systems for DNA mixture profiling. *Theoret. Pop. Biol.*, 63 (2003) 191–205.

[4] I. W. Evett, C. Buffery, G. Wilcott, and D. Stoney. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. *J. Forensic Sci. Soc.*, 31 (1991) 41–47.

[5] B. S. Weir, C. M. Triggs, L. Starling, L. I. Stowell, K. A. J. Walsh, and J. S. Buckleton. Interpreting DNA mixtures. *J. Forensic Sci.*, 42 (1997) 213–222.

[6] T. M. Clayton, J. P. Whitaker, R. Sparkes, and P. Gill. Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Sci. Int.*, 91 (1998) 55–70.

[7] P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton. Interpreting simple STR mixtures using allele peak areas. *Forensic Sci. Int.*, 91 (1998) 41–53.

[8] M. Bill, P. Gill, J. Curran, T. Clayton, R. Pinchin, M. Healy, and J. Buckleton. PENDULUM — a guideline - based approach to the interpretation of STR mixtures. *Forensic Sci. Int.*, 148 (2005) 181–189.

[9] M.W. Perlin and B. Szabady. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *J. Forensic Sci.*, 46 (2001) 1372–1378.

[10] T. Wang, N. Xue, and R. Wickenheiser. Least square deconvolution (LSD): A new way of resolving STR/DNA mixture samples. Presentation at the 13th International Symposium on Human Identification, October 7–10, 2002, Phoenix, AZ, 2002.

[11] I. Evett, P. Gill, and J. Lambert. Taking account of peak areas when interpreting mixed DNA profiles. *J. Forensic Sci.*, 43 (1998) 62–69.

[12] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.

[13] S. L. Lauritzen and F. Jensen. Stable local computation with conditional Gaussian distributions. *Statistics and Computing*, 11 (2001) 191–203.

[14] Y. Torres, I. Flores, V. Prieto, M. Lopez-Soto, M. J. Farfan, A. Carracedo, and P. Sanz. DNA mixtures in forensic casework: a 4-year retrospective study. *Forensic Sci. Int.*, 134 (2003) 180–186.

[15] J. M. Butler, R. Schoske, P. M. Vallone, J. W. Redman, and M. C. Kline. Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American and Hispanic populations. *J. Forensic Sci.*, 48 (2003) 908–911. Available online at www.astm.org.

[16] T. M. Clayton and J. S. Buckleton. Mixtures. In S. J. Walsh J. S. Buckleton, C. M. Triggs, editor, *Forensic DNA Evidence Interpretation*, chapter 7, pages 217–274. CRC Press, 2004.

[17] R. G. Cowell, S. L. Lauritzen, and J. Mortera. Identification and separation of DNA mixtures using peak area information. Statistical Research Paper 25, Sir John Cass Business School, City University London, Nov 2004.

[18] P. Gill, R. Sparkes, and C. Kimpton. Development of guidelines to designate allele using an STR multiplex system. *Forensic Sci. Int.*, 89 (1997) 185–197.

[19] D. Koller and A. Pfeffer. Object-oriented Bayesian networks. In Dan Geiger and Prakash P. Shenoy, editors, *UAI '97: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, August 1-3, 1997, Brown University, Providence, Rhode Island, USA*, pages 302–313. Morgan Kaufmann, 1997.

[20] K. B. Laskey and S. M. Mahoney. Network fragments: Representing knowledge for constructing probabilistic models. In Dan Geiger and Prakash P. Shenoy, editors, *UAI '97: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence,*

August 1-3, 1997, Brown University, Providence, Rhode Island, USA, pages 302–313. Morgan Kaufmann, 1997.

[21] A. P. Dawid. An object-oriented Bayesian network for estimating mutation rates. In Christopher M. Bishop and Brendan J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Jan 3–6 2003, Key West, Florida*, 2003. Available online at: `http://research.microsoft.com/conferences/AIStats2003`.

[22] D. Cavallini and F. Corradi. OOBN for forensic identification through searching a DNA profiles' database. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Jan 6-8, 2005, Savannah Hotel, Barbados*, pages 41–48. Society for Artificial Intelligence and Statistics, 2005. Available online at: `http://www.gatsby.ucl.ac.uk/aistats`.

[23] A. Urquhart, C. P. Kimpton, T. J. Downes, and P. Gill. Variation in short tandem repeat sequences – a survey of twelve microsatellite loci for use as forensic identification markers. *Int. J. Leg. Med.*, 107 (1994) 13–20.

A Likelihoods from peak area information

All Bayesian networks in the present paper have a common structure, as outlined below: The nodes corresponding to the observed relative peak weights $R = (R_a, a = 1, \ldots, A)$ are all continuous. We are interested in the distribution $P(D, M \,|\, R, E)$ where R denotes the peak weight, E other types of evidence, e.g. evidence about genotypes of certain individuals, M the nodes for the mean peak weights, and D the remaining discrete nodes in the network.

The mean peak weights are represented by discrete nodes with possible values $\mu = (\mu_1, \ldots, \mu_A)$. Using Bayes' formula, and the fact that R is conditionally independent of (D, E) given M, it holds that

$$P(D, M \,|\, R, E) \propto P(R \,|\, M) P(D, M, E).$$

Thus, the information in the relative peak weights enter the calculations only through the *peak weight likelihood*

$$L(\mu) = P(R \,|\, M = \mu).$$

In the following we shall find a simple expression for this likelihood.

Consider now a vector $X = (X_1, \ldots, X_A)$ of independent and normally distributed random variables with $X_a \sim \mathcal{N}(\mu_a, \tau_a^2)$, *i.e.* with joint density

$$f(x_1, \ldots, x_d \,|\, \mu, T) = (2\pi)^{-d/2} \prod_{a=1}^{A} \tau_a^{-1} \exp\left\{ -\sum_a \frac{(x_a - \mu_a)^2}{2\tau_a^2} \right\}$$

where we have let $\mu = (\mu_1, \ldots, \mu_A)$ and T a diagonal matrix with τ_a^2 as diagonal elements, *i.e.* $T = \mathrm{diag}(\tau_1^2, \ldots, \tau_A^2)$. We recall that $\sum_a \mu_a = 1$.

The distribution of the sum $S = \sum_a X_a$ is normal $S \sim \mathcal{N}(1, \tau^2)$ with $\tau^2 = \sum_a \tau_a^2$ and the conditional distribution of X given $S = 1$ is itself multivariate normal with the same mean vector μ and covariance matrix T^*, where

$$\tau^*_{aa} = \frac{\tau_a^2(\tau^2 - \tau_a^2)}{\tau^2}, \quad \tau^*_{ab} = \frac{-\tau_a^2\tau_b^2}{\tau^2}.$$

The density of the conditional distribution can now be calculated as

$$f^*(x_1, \ldots, x_{A-1} \,|\, \mu, T^*) \propto \frac{f(x_1, \ldots, x_A \,|\, \mu, T)}{f_S(1 \,|\, \tau^2)} \propto \tau f(x_1, \ldots, x_A \,|\, \mu, T), \tag{4}$$

for $x_A = 1 - \sum_{a=1}^{A-1} x_a$. If we consider the case

$$\tau_a^2 = \sigma^2\mu_a + \omega^2,$$

i.e. the variance structure in (3), we note that

$$\tau^2 = \sum_a \tau_a^2 = \sigma^2 + A\omega^2$$

is constant in μ. Also the covariance matrix T^* is

$$\tau^*_{aa} = \sigma^2\mu_a(1 - \mu_a) + \omega^2 + o(\omega^2), \quad \tau^*_{ab} = -\sigma^2\mu_a\mu_b + o(\omega^2),$$

which, ignoring terms $o(\omega^2)$ which are an order of magnitude smaller than ω^2, has precisely the form (2) used in our model. If we ignore measurement error by setting $\omega^2 = 0$, the entries in T^* are exactly given by (2).

It follows that to an excellent approximation — exact for $\omega^2 = 0$ — *we can calculate the correct peak weight likelihood based on T^* by using the variance structure T with independence:*

$$L(\mu) = f^*(x_1, \ldots, x_{A-1} \,|\, \mu, T^*) \propto f(x_1, \ldots, x_A \,|\, \mu, T),$$

which justifies the use of (3) in the calculations.

B Description of the network classes in the object-oriented network

Below we describe the component networks (together with their internal structure) which are used in the construction of the master network. In what follows, **bold** will indicate a network class, and `teletype` will indicate a node. In the figures, instances of a certain class are represented by a rounded rectangle, discrete nodes have a single outline, whereas continuous nodes have a double outline. Interface nodes are represented with a grey ring; input nodes having a dotted outline and output nodes having a solid outline. Also, dark grey nodes will indicate where possible evidence might be inserted and black nodes are target nodes or nodes of interest where results will be read.

B.1 The founder class

The class **founder** of Figure 10 contains a single node `founder` with the relevant repertory of alleles as its states, and an associated probability table describing their gene frequencies.

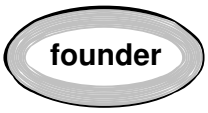


Figure 10: Network **founder** for founder gene.

For illustration, we show marker FGA having observed alleles coded A to C and the aggregation of all unobserved alleles coded as x. The probability table is shown in Table 17.

Table 17: Gene frequencies for marker FGA as reported in Evett *et al.* (1998).

Allele	A	B	C	x
Frequency	0.187	0.165	0.139	0.509

B.2 The genotype class

The class **gt** in Figure 11 represents an individual's genotype `gt`, formed by the unordered pair of paternal and maternal genes, {`pg`, `mg`}. (Input nodes `pg` and `mg` are copies of node `founder` of class **founder**.) The paternal and maternal genes, `pg` and `mg`, are chosen independently from the same population whose allele frequencies are assumed known. Output node `gt` is the logical combination of input nodes `pg` and `mg`.

B.3 The query class

The class **whichgt** of Figure 12 describes the selection between two genotypes.

If the Boolean node `query?` is *true*, then output node, `outgt`, will have identical genotype to `ingt`; otherwise it will be identical to `othergt`. This is written in the HUGIN expression language as: `outgt` := if(`query?` == *true*, `ingt`, `othergt`).[4]

[4]The function if(C, x, y) takes the value x if condition C is satisfied, otherwise y.

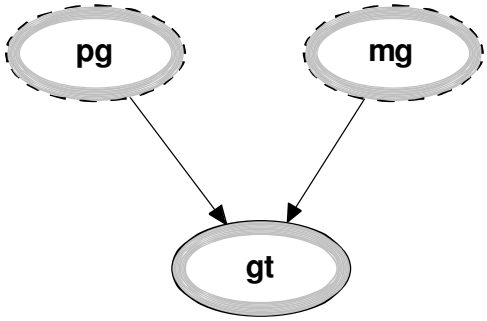


Figure 11: Network **gt** for genotype.

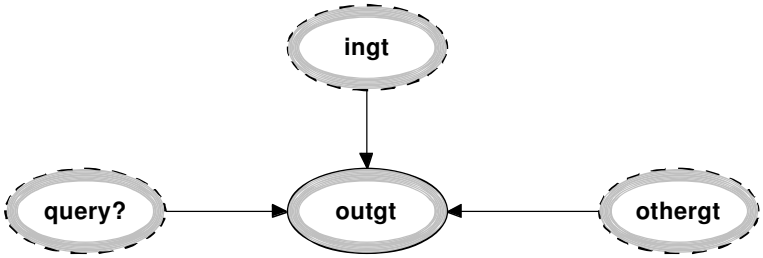


Figure 12: Network **whichgt** for selecting a genotype.

B.4 The joint genotype

The network class **jointgt** of Figure 13 represents the combined genotype of two individuals, p1 and p2. Node `p1gt&p2gt` is simply the logical combination of the two input genotypes in `p1gt` and `p2gt`.

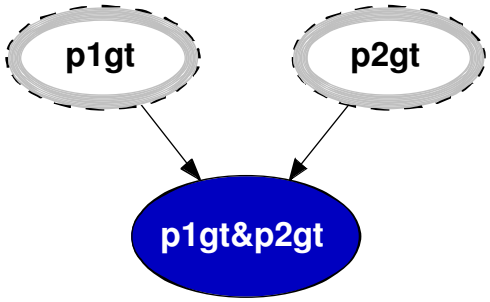


Figure 13: Network **jointgt** for genotype pairs.

B.5 The number of alleles

The class **nalleles** shown in Figure 14 counts the number, varying from 0 to 2, of a certain allelic type in a genotype. For allele A , `nA` := if(`gt` == AA, 2, if (or (`gt` == AB, `gt` == AC, `gt` == Ax), 1, 0)). Similarly, for B, C and x. This class models the $n_a^{(i)}$ variables in (1).

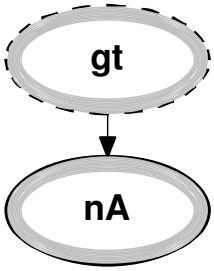


Figure 14: Network **nalleles** for counting the number of alleles.

B.6 The weight of an allele in the mixture

The class **alleleinmix** shown in Figure 15 shows whether a certain allelic type (repeat number) is in the mixture and computes its mean contribution to the peak area of the mixture.

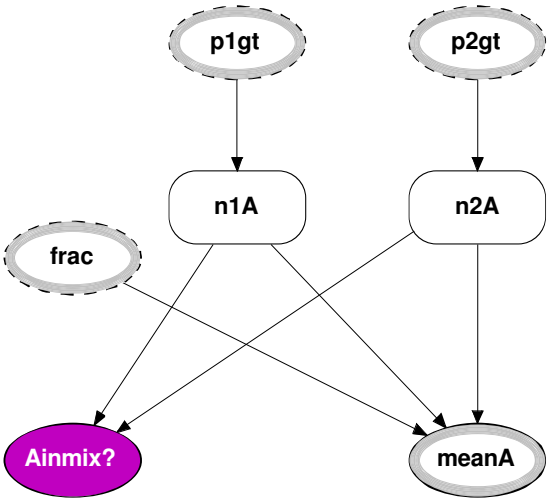


Figure 15: Network **alleleinmix** for alleles in mixture.

Input nodes `p1gt` and `p2gt`, the genotypes of the two people, p1 and p2, contributing to the mixture, have identity links to the input node `gt` in the two instances of class **nalleles**, `n1A` and `n2A`. The Boolean node `Ainmix?` is *true* if at least one of the two contributors has allele A. Thus, `Ainmix?` := if(and (`n1A_nA` == 0, `n2A_nA` == 0), *false*, *true*), where `n1A_nA` and `n2A_nA` refer to the output nodes of the two instances of class **nalleles**, `n1A` and `n2A`. (Similar instances are built for the other alleles.) Repeat number information is entered and propagated from these nodes. For example, if the mixture contains allele A, node `Ainmix?` is set to *true*.

Input node `frac` represents the proportion of DNA contributed by p1, denoted by θ in § 2. To enable evidence propagation in the Bayesian network to be possible, we model this continuous variable by an approximating discrete variable. In our hand-built HUGIN networks we used a coarse level of discretization, with the states of node `frac` put on a discrete scale ranging from $[0, 5]$ for convenience. The scale of node `frac` can easily be modified to a finer grid, though some conditional probability tables dependent on the grid size can get quite large and tedious to fill in correctly by hand if the grid is too fine. Output node `meanA`:= `n1A_nA` * frac + `n2A_nA` * (5 - frac). This differs from the expression for the mean in (1) by a scale factor of 10 which is appropriately corrected for throughout.

B.7 The peak weight

The class **peakweight** shown in Figure 16 models the observable peak weight as in (1).

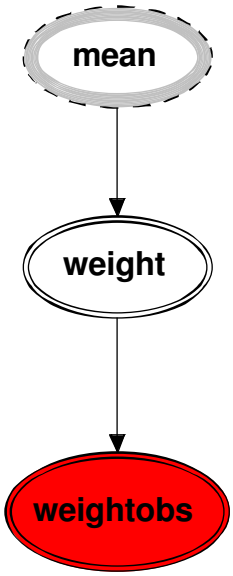


Figure 16: Network **peakweight** for peak weight.

The input node `mean` is identified, for example, with output node `meanA` of class **alleleinmix**. The intermediate continuous node `weight` represents the unobserved true peak weight. This node has a conditional Gaussian distribution with mean given by the value of the discrete parent node `mean` and variance equal to $10 * 0.01 * \texttt{mean}$, representing variations in the amplification process, cf. § 2. The observed peak weight is modelled by the continuous node `weightobs` to allow for additional measurement error of the true weight. When using peak area information the value of the relative peak weight is inserted as evidence in the node `weightobs`.

B.8 The target class

The class **target** shown in Figure 17 has two Boolean output nodes `p1=s?` (`p2=v?`) with *true*, *false* states, representing whether contributor p1 (p2) is the *suspect* (*victim*) or not. The black `target` node is the logical conjunction of the two nodes `p1=s?` and `p2=v?`. These nodes are given a uniform prior distribution so that `target` node has a uniform prior distribution over its states. This enables the computation of the likelihood ratio as described in [3].

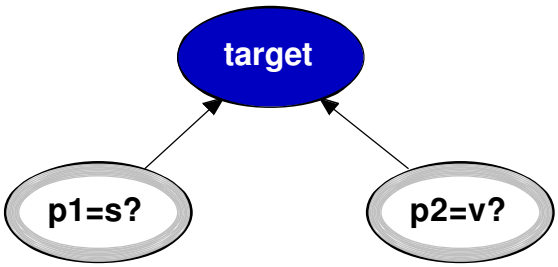


Figure 17: Network **target**.

B.9 The marker class

The class **marker** in Figure 18 is an upper level network containing several instances of the classes defined above. This class is made to represent information related to a particular marker. Here it is illustrated for a marker having three alleles represented in the mixture. Input nodes `spg`, `smg`, `u1pg`, `u1mg`, `vpg`, `vmg`, `u2pg` and `u2mg` are all copies of node `founder` of class **founder**; u1 and u2 being two unspecified individuals. Input nodes `p1=s?` and `p2=v?` are identified with the corresponding output nodes of class **target**. The nodes `sgt`, `u1gt`, `vgt` and `u2gt` are all instances of class **gt**. Evidence on the suspect's and victim's genotypes is entered in the network in the nodes `sgt` and `vgt`. Nodes `p1gt` and `p2gt` are instances of **whichgt** and when `p1=s?` is *true* (*false*), `p1gt` will be identical to `sgt` (`u1gt`). A similar relationship holds between nodes `p2=v?`, `p2gt`, `vgt` and `u2gt`. The node `jointgt` is an instance of **jointgt**; `Amean`, `Bmean`, `Cmean` and `xmean` are instances of **alleleinmix**; `Aweight`, `Bweight`, `Cweight` and `xweight` are instances of **peakweight**. Input node `frac` is identified with the corresponding node in the master network described below.

B.10 The master network

Figure 19 gives the master network used for both identification and separation of DNA mixtures from two contributors. It refers to the data from [11] shown in Table 3.

`D8`, `D18`, `FGA`, and `THO1` are all instances of **marker**; `D21` and `VWA` are instances of a simple modification of class **marker** and the other network classes it calls, in order to account for 4 observed alleles. `D8`, `D18`, `FGA`, `THO1`, `D21` and `VWA` each have 8 **founder** instances with their appropriate gene frequencies as input to the 8 input nodes of class **marker**. The `frac` node is connected to all the markers showing their dependence via this quantity. `target`, an instance of class **target**, is linked to each marker via its output nodes `p1=s?` and `p2=v?`. Once constructed, the master network can be used to insert and propagate case evidence in the appropriate internal nodes, and the marginal posterior probability distributions of the quantities of interest can be read from the corresponding nodes.

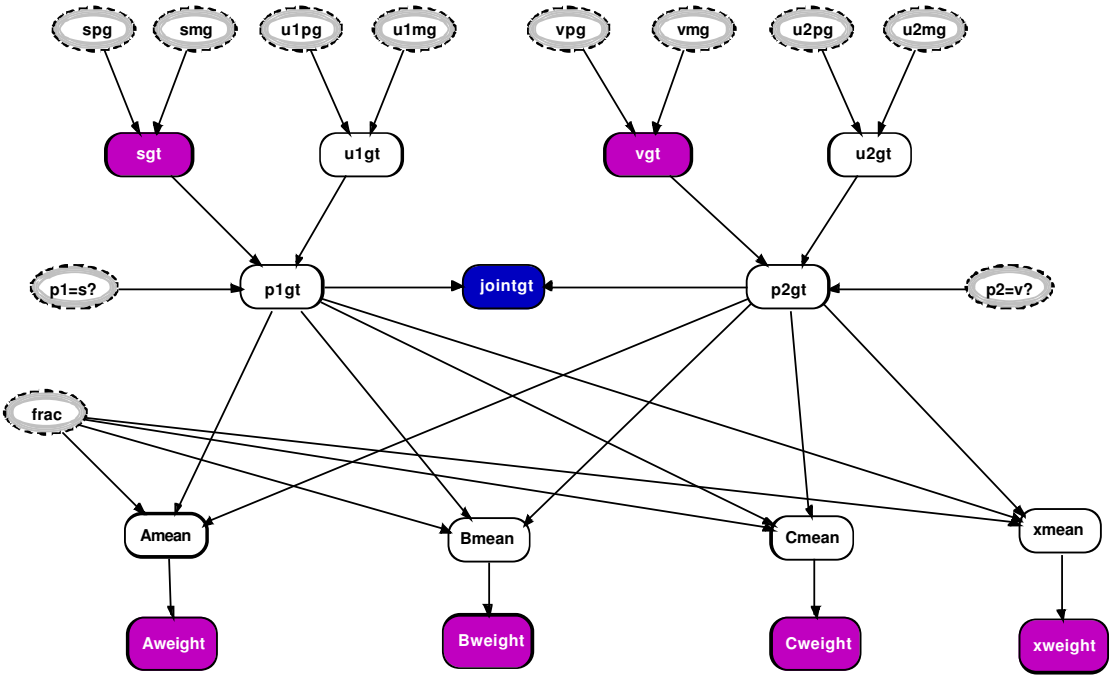


Figure 18: Network **marker** with three observed allele peaks.

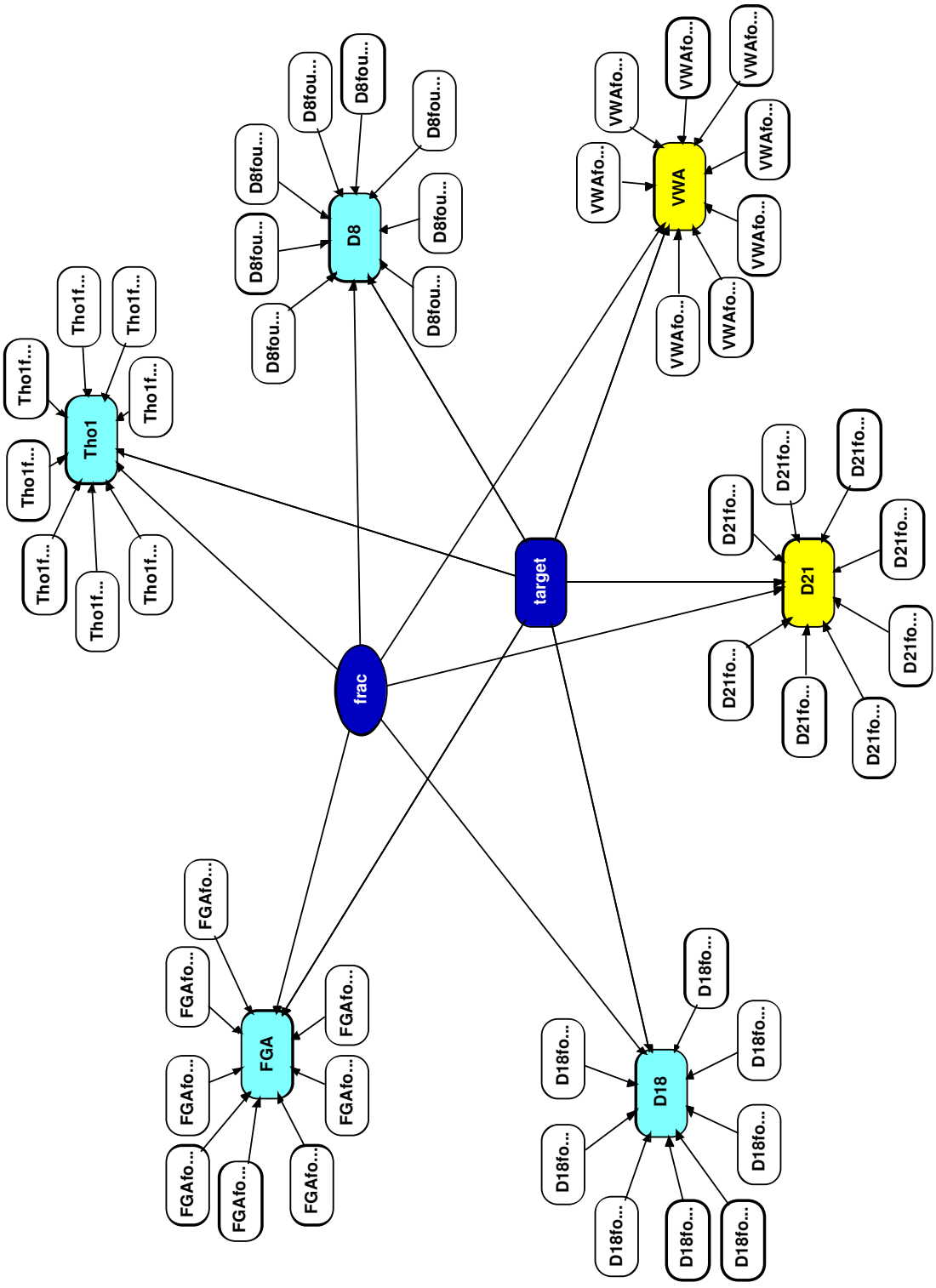


Figure 19: Master network for identification and separation of mixtures.

B.11 Amelogenin marker

To build a network for amelogenin one needs to make the following changes to the previous classes. No founder class is needed and the genotype class has a single output node `gt` with states XX for female and XY for male, with equal prior probabilities. The **query** and **jointgt** classes only need trivial modifications to reduce their state spaces. The allele counting class **nalleles**, for a male contributor, `gt`== XY, (for a female contributor, `gt`== XX) has `nX`==1 (2) and `nY`== 1 (0). The class **alleleinmix** of § B.6, is modified so that `Xinmix?` is set to *true*. All other network classes remain unchanged.

C Description of the networks generated by Maies

The Bayesian network generated by Maies may be considered equivalent to an "unfolded" version of the object-oriented networks described in Appendix B. An example of a network generated for a single marker with two alleles observed in the mixture is shown in Figure 20. The structure is similar to the network shown in Figure 18, and like the object-oriented network described earlier there are several distinct modules of repetition that can be seen in the figure: indeed it is this repetitive structure that makes it possible for Maies to create the much larger Bayesian networks required to analyse mixtures on several markers. We now describe these various structures and how they interrelate.

C.1 Founding people

Maies currently assumes that DNA from two individuals are in the mixture. Thus it sets up nodes for four founding individuals who are paired up, prefixed by `s` (for suspect), `v` (for victim), and `u1` and `u2` representing two unspecified persons from the population. Corresponding to each of these individuals is a triple of nodes representing their genotype on the marker, and the individuals' paternal and maternal genes. They are joined up as in Figure 11 and their function is the same. The probability tables associated with the maternal and paternal genes contain the allele frequencies of the observed alleles, whilst the conditional probability table associated with the genotype node is the logical combination of the maternal and paternal gene.

C.2 Actual contributors to the mixture

The genotypes on the marker of the two individuals p1 and p2 whose DNA is in the mixture are the nodes labelled `p1gt` and `p2gt`. Node `p1gt` has incoming arrows from nodes `u1gt`, `sgt` and a `(yes,no)` valued binary node labelled `p1 = s?`. The function of this latter node is similar to the `query?` node of Figure 12, namely to set the genotype of node `p1gt` to be that of `sgt` if `p1 = s?` takes the value `yes`, otherwise set the genotype of node `p1gt` to be that of `u1gt`. An equivalent relationship holds between the genotype nodes `p2gt`, `vgt`, `u2gt` and `p2 = v?`.

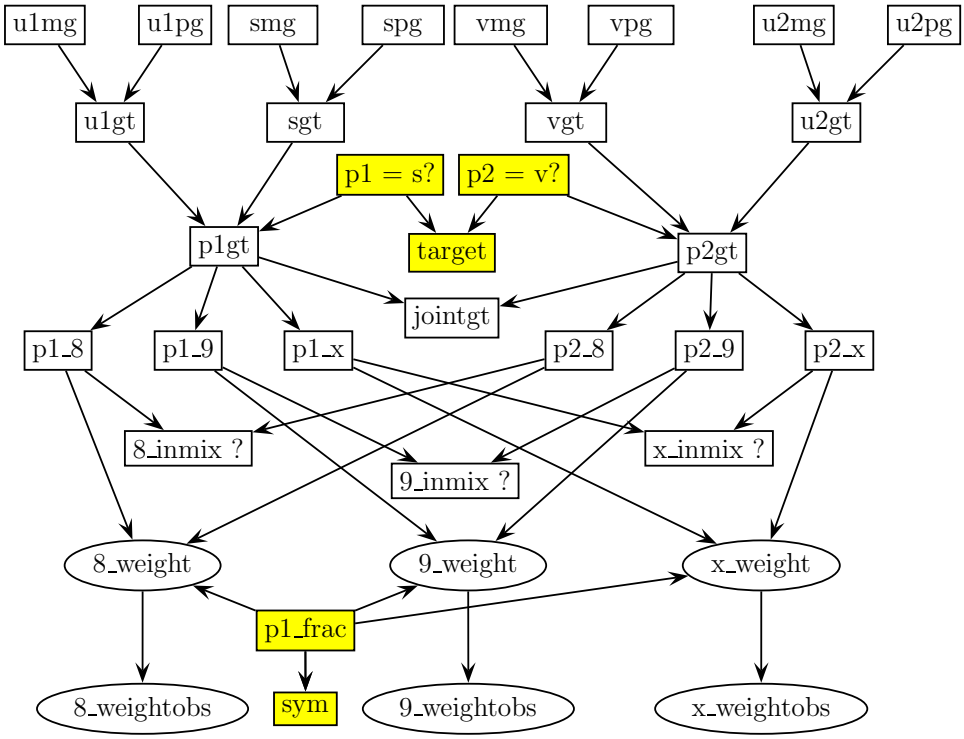


Figure 20: The structure of a Bayesian network generated by MAIES for a single marker, in which two allele peaks (8 and 9) were observed.

The node labelled `target` represents the four possible combinations of values of the two nodes `p1 = s?` and `p2 = v?` as in Figure 17 and described in § B.8.

The network also has a node representing the joint genotypes of individuals p1 and p2, which is labelled `jointgt`, with incoming arrows from `p1gt` and `p2gt`; the function of this part of the network is equivalent to the object shown in Figure 13.

C.3 Allele counting nodes

On the level below the genotype nodes for p1 and p2 is a set of nodes representing the number of alleles (taking the value of 0, 1 or 2) of a certain type in each individual. Thus, for example, the node `p1_8` counts the number of alleles of repeat number 8 in the genotype of individual p1 for the given marker: this value only depends upon the genotype of the individual p1 and hence there is an arrow from `p1gt` to `p1_8`. These nodes model the $n_a^{(i)}$ variables introduced in (1).

C.4 Repeat number nodes

On the level below the allele counting nodes are the repeat number nodes, labelled `8_inmix?`, `9_inmix?` and `x_inmix?`. These are `(yes,no)` binary valued nodes representing whether or not the particular alleles are present in the mixture: thus for example allele 8 is present in the mixture if either of the allele counting nodes `p1_8` or `p2_8` takes a non-zero value. For the node `x_inmix?` the `x` refers to all of the alleles in the marker that are not observed. When using repeat number information as evidence the repeat number nodes present in the mixture will be given the value `yes`; all other nodes, including `x_inmix?`, will be given the value `no`.

C.5 True and observed weight nodes

These nodes are represented by the rounded rectangle shapes. The nodes `8_weight`, `9_weight` and `x_weight` represent the true relative peak weights r_8, r_9 and r_x respectively of the alleles 8, 9 and `x` in the amplified DNA sample; the nodes `8_weightobs`, `9_weightobs` and `x_weightobs` represent the measured weights. The observed weight is given a conditional-Gaussian distribution with mean the true weight, and variance ω^2. Each true-weight node is given a conditional-Gaussian distribution with mean $\mu_a = \{\theta n_a^{(1)} + (1-\theta) n_a^{(2)}\}/2$, where the fraction θ of DNA from p1 in the mixture is modelled in the network by a discrete distribution in the node labelled `p1_frac`. The variance is taken to be $\sigma^2 \mu_a$, as specified in § 2.

When using peak area information as evidence the nodes representing the observed weights will have their values set to the relative peak weights. The `sym` node is only used for separating a mixture of two unknown contributors, as described in § 5.2.

C.6 Networks with more than one marker

The network displayed in Figure 20 generated by MAIES is for a single marker; for mixture problems involving several markers the structure is similar but more complex because the number of nodes grows with the number of markers (in the *Graham* example, see § 4.1, there are 325 nodes). In such a network the nodes shaded in Figure 20 occur only once. The unshaded nodes are replicated once for each marker, with each node having text in their labels to identify the marker that the allele or genotype nodes refer to. There will also be extra repeat number, allele counting and allele weight nodes in each marker having more than two observed alleles in the mixture, extending the pattern for the one-marker network in the obvious manner.

Cass Business School
City of London

FACULTY OF ACTUARIAL SCIENCE AND STATISTICS

Actuarial Research Papers since 2001

Report Number	Date	Publication Title	Author
135.	February 2001.	On the Forecasting of Mortality Reduction Factors. ISBN 1 901615 56 1	Steven Haberman Arthur E. Renshaw
136.	February 2001.	Multiple State Models, Simulation and Insurer Insolvency. ISBN 1 901615 57 X	Steve Haberman Zoltan Butt Ben Rickayzen
137.	September 2001	A Cash-Flow Approach to Pension Funding. ISBN 1 901615 58 8	M. Zaki Khorasanee
138.	November 2001	Addendum to "Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving". ISBN 1 901615 59 6	Peter D. England
139.	November 2001	A Bayesian Generalised Linear Model for the Bornhuetter-Ferguson Method of Claims Reserving. ISBN 1 901615 62 6	Richard J. Verrall
140.	January 2002	Lee-Carter Mortality Forecasting, a Parallel GLM Approach, England and Wales Mortality Projections. ISBN 1 901615 63 4	Arthur E.Renshaw Steven Haberman.
141.	January 2002	Valuation of Guaranteed Annuity Conversion Options. ISBN 1 901615 64 2	Laura Ballotta Steven Haberman
142.	April 2002	Application of Frailty-Based Mortality Models to Insurance Data. ISBN 1 901615 65 0	Zoltan Butt Steven Haberman
143.	Available 2003	Optimal Premium Pricing in Motor Insurance: A Discrete Approximation.	Russell J. Gerrard Celia Glass
144.	December 2002	The Neighbourhood Health Economy. A Systematic Approach to the Examination of Health and Social Risks at Neighbourhood Level. ISBN 1 901615 66 9	Les Mayhew
145.	January 2003	The Fair Valuation Problem of Guaranteed Annuity Options : The Stochastic Mortality Environment Case. ISBN 1 901615 67 7	Laura Ballotta Steven Haberman
146.	February 2003	Modelling and Valuation of Guarantees in With-Profit and Unitised With-Profit Life Insurance Contracts. ISBN 1 901615 68 5	Steven Haberman Laura Ballotta Nan Want
147.	March 2003.	Optimal Retention Levels, Given the Joint Survival of Cedent and Reinsurer. ISBN 1 901615 69 3	Z. G. Ignatov Z.G., V.Kaishev R.S. Krachunov
148.	March 2003.	Efficient Asset Valuation Methods for Pension Plans. ISBN 1 901615707	M. Iqbal Owadally
149.	March 2003	Pension Funding and the Actuarial Assumption Concerning Investment Returns. ISBN 1 901615 71 5	M. Iqbal Owadally

150.	Available August 2004	Finite time Ruin Probabilities for Continuous Claims Severities	D. Dimitrova Z. Ignatov V. Kaishev
151.	August 2004	Application of Stochastic Methods in the Valuation of Social Security Pension Schemes. ISBN 1 901615 72 3	Subramaniam Iyer
152.	October 2003.	Guarantees in with-profit and Unitized with profit Life Insurance Contracts; Fair Valuation Problem in Presence of the Default Option[1]. ISBN 1-901615-73-1	Laura Ballotta Steven Haberman Nan Wang
153.	December 2003	Lee-Carter Mortality Forecasting Incorporating Bivariate Time Series. ISBN 1-901615-75-8	Arthur E. Renshaw Steven Haberman
154.	March 2004.	Operational Risk with Bayesian Networks Modelling. ISBN 1-901615-76-6	Robert G. Cowell Yuen Y, Khuen Richard J. Verrall
155.	March 2004.	The Income Drawdown Option: Quadratic Loss. ISBN 1 901615 7 4	Russell Gerrard Steven Haberman Bjorn Hojgarrd Elena Vigna
156.	April 2004	An International Comparison of Long-Term Care Arrangements. An Investigation into the Equity, Efficiency and sustainability of the Long-Term Care Systems in Germany, Japan, Sweden, the United Kingdom and the United States. ISBN 1 901615 78 2	Martin Karlsson Les Mayhew Robert Plumb Ben D. Rickayzen
157.	June 2004	Alternative Framework for the Fair Valuation of Participating Life Insurance Contracts. ISBN 1 901615-79-0	Laura Ballotta
158.	July 2004.	An Asset Allocation Strategy for a Risk Reserve considering both Risk and Profit. ISBN 1 901615-80-4	Nan Wang
159.	December 2004	Upper and Lower Bounds of Present Value Distributions of Life Insurance Contracts with Disability Related Benefits. ISBN 1 901615-83-9	Jaap Spreeuw
160.	January 2005	Mortality Reduction Factors Incorporating Cohort Effects. ISBN 1 90161584 7	Arthur E. Renshaw Steven Haberman
161.	February 2005	The Management of De-Cumulation Risks in a Defined Contribution Environment. ISBN 1 901615 85 5.	Russell J. Gerrard Steven Haberman Elena Vigna
162.	May 2005	The IASB Insurance Project for Life Insurance Contracts: Impart on Reserving Methods and Solvency Requirements. ISBN 1-901615 86 3.	Laura Ballotta Giorgia Esposito Steven Haberman
163.	September 2005	Asymptotic and Numerical Analysis of the Optimal Investment Strategy for an Insurer. ISBN 1-901615-88-X	Paul Emms Steven Haberman
164.	October 2005.	Modelling the Joint Distribution of Competing Risks Survival Times using Copula Functions. I SBN 1-901615-89-8	Vladimir Kaishev Dimitrina S, Dimitrova Steven Haberman
165.	November 2005.	Excess of Loss Reinsurance Under Joint Survival Optimality. ISBN1-901615-90-1	Vladimir K. Kaishev Dimitrina S. Dimitrova
166.	November 2005.	Lee-Carter Goes Risk-Neutral. An Application to the Italian Annuity Market. ISBN 1-901615-91-X	Enrico Biffis Michel Denuit

167.	November 2005	Lee-Carter Mortality Forecasting: Application to the Italian Population. ISBN 1-901615-93-6	Steven Haberman Maria Russolillo

Statistical Research Papers

Report Number	Date	Publication Title	Author
1.	December 1995.	Some Results on the Derivatives of Matrix Functions. ISBN 1 874 770 83 2	P. Sebastiani
2.	March 1996	Coherent Criteria for Optimal Experimental Design. ISBN 1 874 770 86 7	A.P. Dawid P. Sebastiani
3.	March 1996	Maximum Entropy Sampling and Optimal Bayesian Experimental Design. ISBN 1 874 770 87 5	P. Sebastiani H.P. Wynn
4.	May 1996	A Note on D-optimal Designs for a Logistic Regression Model. ISBN 1 874 770 92 1	P. Sebastiani R. Settimi
5.	August 1996	First-order Optimal Designs for Non Linear Models. ISBN 1 874 770 95 6	P. Sebastiani R. Settimi
6.	September 1996	A Business Process Approach to Maintenance: Measurement, Decision and Control. ISBN 1 874 770 96 4	Martin J. Newby
7.	September 1996.	Moments and Generating Functions for the Absorption Distribution and its Negative Binomial Analogue. ISBN 1 874 770 97 2	Martin J. Newby
8.	November 1996.	Mixture Reduction via Predictive Scores. ISBN 1 874 770 98 0	Robert G. Cowell.
9.	March 1997.	Robust Parameter Learning in Bayesian Networks with Missing Data. ISBN 1 901615 00 6	P.Sebastiani M. Ramoni
10.	March 1997.	Guidelines for Corrective Replacement Based on Low Stochastic Structure Assumptions. ISBN 1 901615 01 4.	M.J. Newby F.P.A. Coolen
11.	March 1997	Approximations for the Absorption Distribution and its Negative Binomial Analogue. ISBN 1 901615 02 2	Martin J. Newby
12.	June 1997	The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases. ISBN 1 901615 10 3	M. Ramoni P. Sebastiani
13.	June 1997	Learning Bayesian Networks from Incomplete Databases. ISBN 1 901615 11 1	M. Ramoni P.Sebastiani
14.	June 1997	Risk Based Optimal Designs. ISBN 1 901615 13 8	P.Sebastiani H.P. Wynn
15.	June 1997.	Sampling without Replacement in Junction Trees. ISBN 1 901615 14 6	Robert G. Cowell
16.	July 1997	Optimal Overhaul Intervals with Imperfect Inspection and Repair. ISBN 1 901615 15 4	Richard A. Dagg Martin J. Newby
17.	October 1997	Bayesian Experimental Design and Shannon Information. ISBN 1 901615 17 0	P. Sebastiani. H.P. Wynn
18.	November 1997.	A Characterisation of Phase Type Distributions. ISBN 1 901615 18 9	Linda C. Wolstenholme
19.	December 1997	A Comparison of Models for Probability of Detection (POD) Curves. ISBN 1 901615 21 9	Wolstenholme L.C
20.	February 1999.	Parameter Learning from Incomplete Data Using Maximum Entropy I: Principles. ISBN 1 901615 37 5	Robert G. Cowell

21.	November 1999	Parameter Learning from Incomplete Data Using Maximum Entropy II: Application to Bayesian Networks. ISBN 1 901615 40 5	Robert G. Cowell
22.	March 2001	FINEX : Forensic Identification by Network Expert Systems. ISBN 1 901615 60X	Robert G.Cowell
23.	March 2001.	Wren Learning Bayesian Networks from Data, using Conditional Independence Tests is Equivalant to a Scoring Metric ISBN 1 901615 61 8	Robert G Cowell
24.	August 2004	Automatic, Computer Aided Geometric Design of Free-Knot, Regression Splines. ISBN 1-901615-81-2	Vladimir K Kaishev, Dimitrina S.Dimitrova, Steven Haberman Richard J. Verrall
25.	December 2004	Identification and Separation of DNA Mixtures Using Peak Area Information. ISBN 1-901615-82-0	R.G.Cowell S.L.Lauritzen J Mortera,
26.	November 2005.	The Quest for a Donor : Probability Based Methods Offer Help. ISBN 1-90161592-8	P.F.Mostad T. Egeland., R.G. Cowell V. Bosnes Ø. Braaten
27.	February 2006	Identification and Separation of DNA Mixtures Using Peak Area Information. (Updated Version of Research Report Number 25). ISBN 1-901615-94-4	R.G.Cowell S.L.Lauritzen J Mortera,

Faculty of Actuarial Science and Statistics

Actuarial Research Club

The support of the corporate members

CGNU Assurance
English Matthews Brockman
Government Actuary's Department

is gratefully acknowledged.

ISBN 1-901615-94-4